AD-A113 778 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
POST-CONFIGURAL POLYSAMPLING PUSHBACK PERFORMANCE.(U)
FEB 82 K B KRYSTINIK DAAG29-79-C-0205
UNCLASSIFIED ARO-16669.14-M NL

1 of 1
AD A
113778

END
DATE
FILMED
5-82
DTIC

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(12)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: ARO 16669.14-M	2. GOVT ACCESSION NO.: AD-A113 778	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Post-configural Polysampling Pushback Performance		5. TYPE OF REPORT & PERIOD COVERED: Technical Rept.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Katherine Bell Krystinik		8. CONTRACT OR GRANT NUMBER(s): DAAG29 79 C 0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Princeton University, Princeton, NJ 08540		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709		12. REPORT DATE: 11 Feb 82
		13. NUMBER OF PAGES: 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DTIC ELECTE APR 26 1982 D

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

sampling(statistics)
robust analysis
estimating
statistics

Copy available to DTIC does not permit fully legible reproduction

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The pushback estimates are a preliminary data modification followed by the application of a robust statistic to the modified data. Application of configural polysampling techniques and use of the minimum attainable variance and maximum attainable polyefficiency derived from these techniques aid in fine-tuning the pushback estimates. The form of the pushback estimate shown by traditional Monte Carlo methods to perform well in comparison to a good biweight is modified and the performance is improved.

AD A113778

DTIC FILE COPY

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DISCLAIMER NOTICE

Post-configural polysampling pushback performance*

by

Katherine Bell Krystinik

Technical Report No. 211, Series 2
Department of Statistics
Princeton University
February 1982

*Prepared in connection with research at Princeton University, supported by the Army Research Office (Durham).

ABSTRACT

The pushback estimates are a preliminary data modification followed by the application of a robust statistic to the modified data. Application of configural polysampling techniques and use of the minimum attainable variance and maximum attainable polyefficiency derived from these techniques aid in fine-tuning the pushback estimates. The form of the pushback estimate shown by traditional Monte Carlo methods to perform well in comparison to a good biweight is modified and the performance is improved.

1. Introduction.

The techniques and uses of configural sampling and configural polysampling were described in ((Bell and Pregibon (1981)), (Pregibon and Tukey (1981)) and (Tukey (1981))). To briefly review, the uses are:

(1) the determination of the minimum attainable variance for each sampling situation,

(2) the determination of the maximum attainable polyefficiency for several sampling situations and

(3) the tuning of a robust procedure with the aim of increasing its efficiency or polyefficiency.

(1) can be achieved using configural sampling or polysampling methods. In the former, no weights are used since the data are all from the situation under consideration. In the latter, weights (as described in (Pregibon and Tukey (1981))) are used to take into account the fact that we have data from situations other than that for which we are determining the minimum variance. The results discussed here are based on the configural polysampling techniques (i.e. the weighted case).

These uses of configural polysampling are applied to the pushback procedures. The pushback estimates are defined

*Prepared in connection with research at Princeton University, supported by the Army Research Office (Durham).

as follows: Suppose we are given n observations,

$$y_1, y_2, \ldots, y_n ,$$

from a particular situation $\{f_i: i=1, \ldots, n\}$ where the f_i are location scale densities. The pushback procedure modifies the order statistics of the n observations,

$$y(1), y(2), \ldots, y(n) ,$$

by substracting some function of i, p(i),

$$y(1)-p(1), y(2)-p(2), \ldots, y(n)-p(n) .$$

The form of p(i) considered is

$$p(i) = k \cdot s \cdot a(i)$$

where k is a constant, s is an estimate of the scale of the $\{y(i)\}$ and $\{a(i)\}$ is a set of central values of order statistics from a suitable unit distribution. Application of a robust estimate to the pushed-back data determines the pushback location estimate for the distribution of the $\{y(i)\}$.

2. Minimum attainable variances.

As seen in (Krystinik (1981b)), traditional Monte Carlo results indicate that the pushback estimates of the form P%AD-Gaus-pushback median perform well when maximin efficiencies (with respect to the w6-biweight) are used as the criterion of performance. We check this conclusion using

the minimum variance estimates for the bi-situation, the Gaussian and the slash. Using 300 data configurations, 150 from the Gaussian and 150 from the slash, we obtain minimum variance results as follows (see (Krystinik (1981a)) for the method of calculation):

	minimum variance
Gaussian	.0528
slash	.2534

Although standard error measurements for the variance estimates are still in the rough formulation stage, we note that these estimates should be fairly well determined since, in using configural polysampling, we are effectively getting information on these estimates from many more samples (than configurations). The estimate for each configuration and the variance estimate associated with it contain information for the many samples (r and s varying; see (Tukey and Pregibon (1981))) associated with the data configuration {c(i)}.

Since the minimum variance for the Gaussian is known to be .05 we will use this value and the slash minimum variance value given above to calculate efficiencies for the P%AD-Gaus-pushback median. These are shown in table 1 for a range of P from 37.5 to 75. These efficiences are calculated using the traditional Monte Carlo variances.

From table 1, we see that the maximin efficiency is approximately 76% and is achieved at P=55 for k=0.9.

Table 1

Efficiencies* of the P%AD-Gaus-pushback median
for the Gaussian and slash

	k	P = 37.5	45	50**	55	70	75
slash	.4	.766	.768	--	.773	.759	.716
	.8	.775	.770	.777	.780	.544	.410
	1.0	.782	.756	.752	.747	.416	.344
	1.2	.735	.716	.678	.660	.373	.297
Gaussian	.4	.687	.689	--	.693	.714	.728
	.8	.697	.711	.742	.758	.890	.933
	1.0	.712	.746	.804	.826	.947	.956
	1.2	.733	.795	.870	.891	.945	.921

*with respect to the configural polysampling based minimum variance for the slash and .05 for the Gaussian minimum variance

**50%AD values are those of the MAD.

Thus using an estimate of the actual minimum attainable variance for the slash for sample size 20 (rather than a best known for which w6-biweight was a close approximation, or an asymptotic lower bound) and the two situations (Gaussian and slash) which are likely to cover the remaining 3(OWG, mix and slacu), we obtain conclusions which support those obtained using the w6-biweight variances and the five situations. We limit further discussion to the 55%AD-Gaus-pushback median form.

3. Maximum attainable biefficiency.

Following the computations discussed in (Krystinik (1981a)), we obtain the biefficiency for the two situations, i.e. $\max_t \left| \min_{Q=G,s} \frac{\text{minvar}_Q}{\text{var}_Q(t)} \right|$. The biefficiency for sample size 20 is 96%. The bioptimal curve corresponding to different shadow prices (see (Krystinik and Morgenthaler (1981))) for the two situations is shown in Figure 1. This optimal efficiency can be used to see how far from the optimum possible value a specific robust procedure is. For example, the pushback (55%AD-Gaus-pushback median) biefficiency is 76%. The pushback is doing reasonably well but some fine-tuning to increase its efficiency would be desirable.

4. Fine-tuning the pushback.

The third use of configural polysampling, i.e. fine-tuning robust estimates, here the pushback, is done as follows. Using t=55%AD-Gaus-pushback median, we calculate t

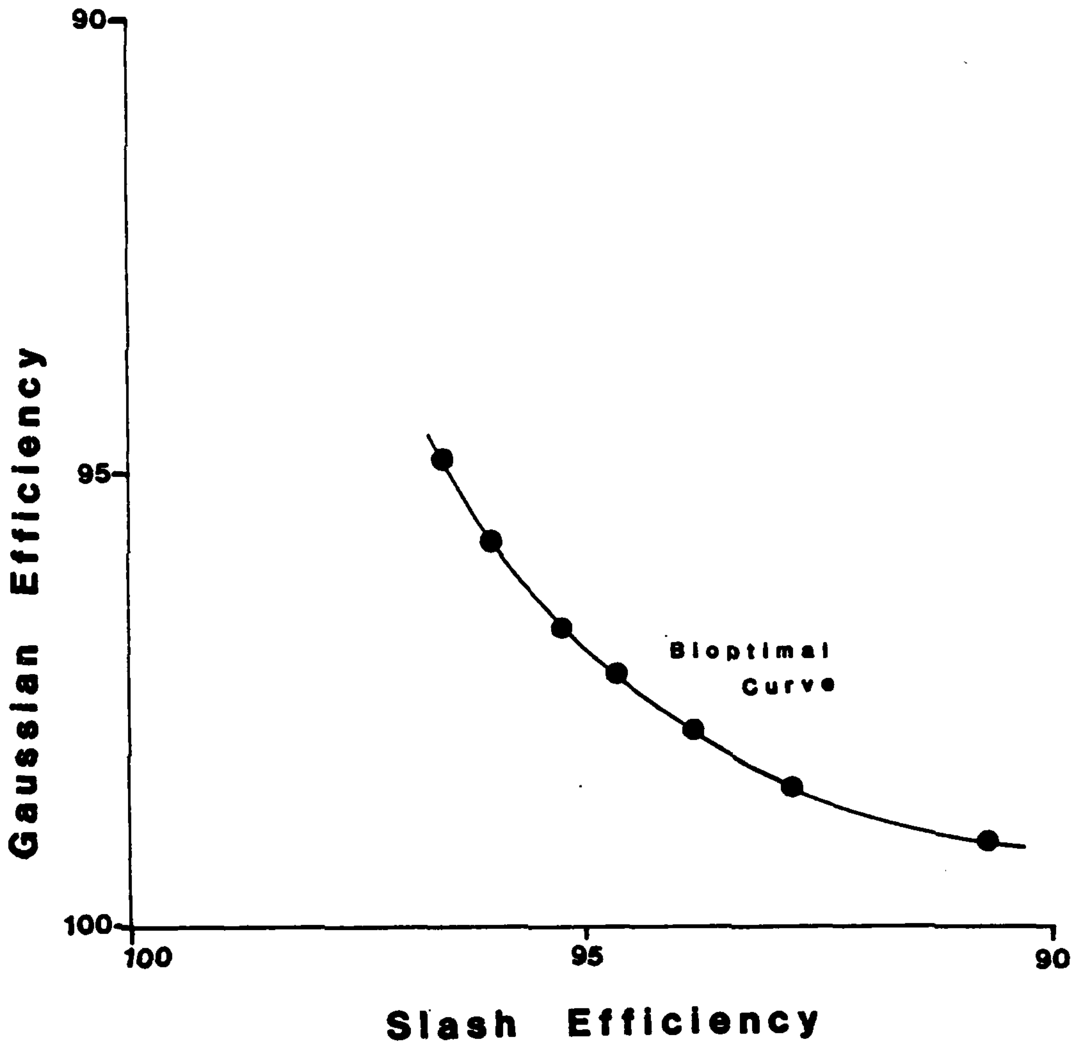

Figure 1: Bioptimal (Gaussian and Slash) curve for sample size twenty.

for each of the 100/150 configurations from the Gaussian and the slash. We then compare the pushback estimate for each configuration with the biefficient estimate. Configurations which exhibit a large difference between these two are noted. We determine whether some slight modification of the pushback procedure will bring the pushback estimate closer to the biefficient estimate in these configurations. The data sample shown in Figure 2 (as a sample, not on the configuration scale; the configuration is just a rescaling and translation of the values) is an example of a configuration for which the pushback estimate and the biefficient estimate are quite different. Note also that the w6-biweight is between the two. Figure 2 shows the original sample with order statistics labelled A-T. The pushback data are shown for k=.8, .9, 1.0, 1.1, and 1.2 on the five lines at the bottom of the figure. Straight lines connect the original order statistics to the associated pushback values. The bioptimal estimate is shown as $\downarrow$ on the figure, the 55%AD-Gaus-pushback median as ¦ , and the w6-biweight as $\curlywedge$.

A modification which eliminates $m \leq 20$ observations (where m depends on the configuration) far from the center of the data and then uses the set $\{a(i)\}$, $i=1, 20-m\}$, the central order-statistic values for a Gaussian sample of size $20-m$, is suggested. This modification tends to keep central Gaussian-like points and uses a set of central order-statistic values adapted to the new sample size. One form of this modification that has been shown to perform well

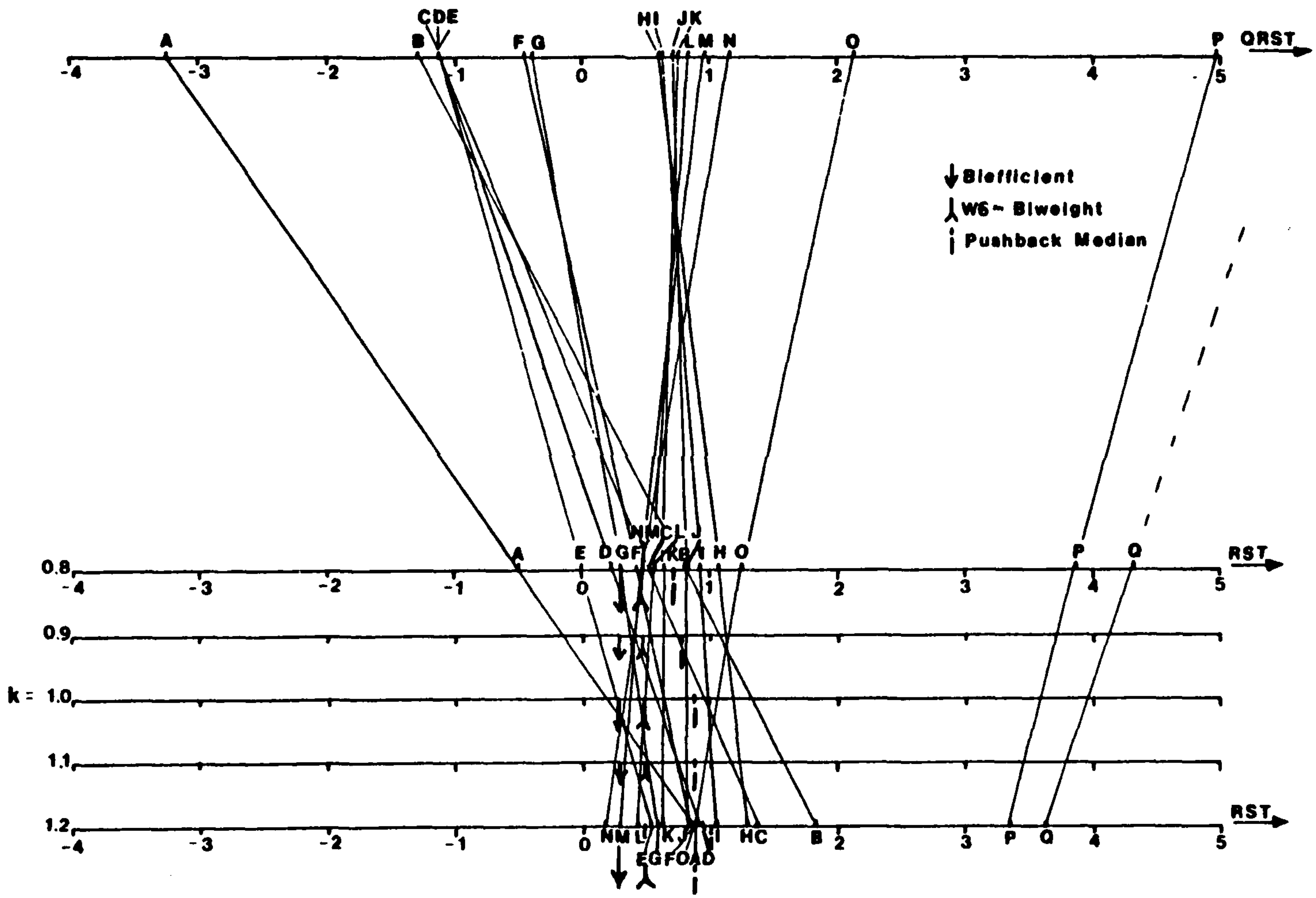

Figure 2: Original and pushback data for a sample where the pushback is quite different from the biefficient estimate. (The biefficient estimate and w6-biweight are for the original data).

(see Andrews et al (1972)) is the set of skipping procedures. Skipping at 1.0 (1.5, 2.0) is defined as follows:

(1) calculate the hinges and the hingespread,

(2) eliminate observations further out than 1.0 (1.5, 2.0) times the hingespread from either hinge.

The skipping procedures were tested with skipping at 1.0, 1.5, and 2.0. Figure 3 shows the application of skipping at 2.0 to the data shown in Figure 2. The skipped pushback estimate has moved closer to the biefficient estimate and is closer to the biefficient estimate than the w6-biweight for pushback constants k=.8, .9, 1.0, and 1.1.

The overall performance of the skipped procedures needs to be evaluated. The skipping modification may improve the performance of the pushback for the configurations on which the pushback and biefficient estimates are far apart, but at the same time make the pushback estimates worse on the other configurations. Table 2 shows the efficiencies (w.r.t. the minimum variance in each situation) of the skipped 55%AD-Gaus-pushback median. These efficiencies are calculated by obtaining variance estimates for the skipped 55%AD-Gaus-pushback median. Skipped 55%AD-Gaus-pushback medians are calculated for the same 150/150 configurations used in obtaining the minimum variance estimates. We then use the relation

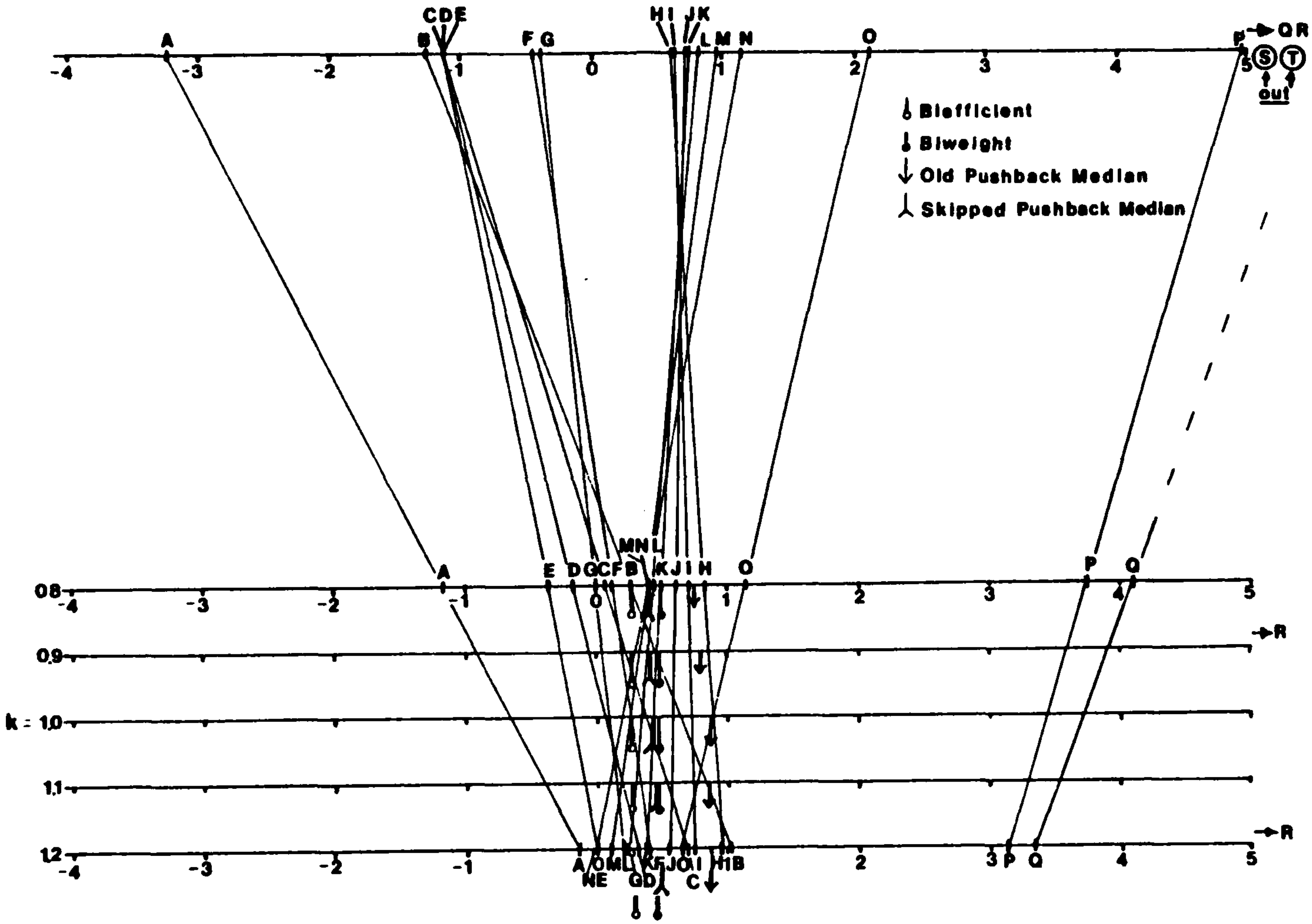

Figure 3: Original and skipped pushback data for the sample of figure 35. (The biefficient estimate and w6-biweight are for the original data).

Table 2

Efficiencies* of the skipped MAD-Gaus-pushback median
for the Gaussian and slash

	k	skipped at 2.0	1.5	1.0
slash	.8	.930	.931	.927
	.9	.933	.936	.925
	1.0	.926	.932	.920
	1.1	.894	.909	.903
	1.2	.875	.890	.884
Gaussian	.8	.747	.732	.675
	.9	.783	.766	.703
	1.0	.825	.805	.731
	1.1	.865	.841	.753
	1.2	.892	.866	.769

*With respect to configural polysampling based minimum variances.

$$MSE\{t(\underline{y})|\underline{c}\} = MSE\{t_o(\underline{y})|\underline{c}\} + E\{s^2|\underline{c}\}(t(\underline{c})-t_o(\underline{c}))^2$$

where $t_o(c)$ is the minimum variance estimate for the configuration and $\{t_o(\underline{y})|\underline{c})\}$ is the rescaled and translated version, $t_o(\underline{y}) = r_{obs}+s_{obs}\cdot t_o(\underline{c})$. Combining the configuration level information $E\{s^2|\underline{c}\}$ with the optimal estimate values and the skipped pushback estimate value, we obtain MSE(t) for a given configuration. We then use the weights described in (Pregibon, Tukey (1981)) to obtain an estimate of the unconditional MSE. As seen in table 2 the biefficiency has increased from 75% to 87.5% due to the configural polysampling guided modification of the pushback.

5. Bioptimal curves and possible further modifications of the pushback.

Figure 4 shows the bioptimal curve and the skipped pushback curves for fixed skipping constants and those for fixed pushback constant. It also shows the bioptimal one-step biweight. For sample size 20, S. Morgenthaler (personal communication, 1981) has shown the best one-step biweight to be the w6.75-biweight. It has a biefficiency of 87%. Thus simple estimates in the form of skipped pushbacks perform very well in comparison to the maximum attainable biefficiency and the w6.75-biweight.

What does this picture (figure 4) suggest for better choices of estimates aimed at achieving 1) higher biefficiency, 2) high slash efficiency with 90% Gaussian

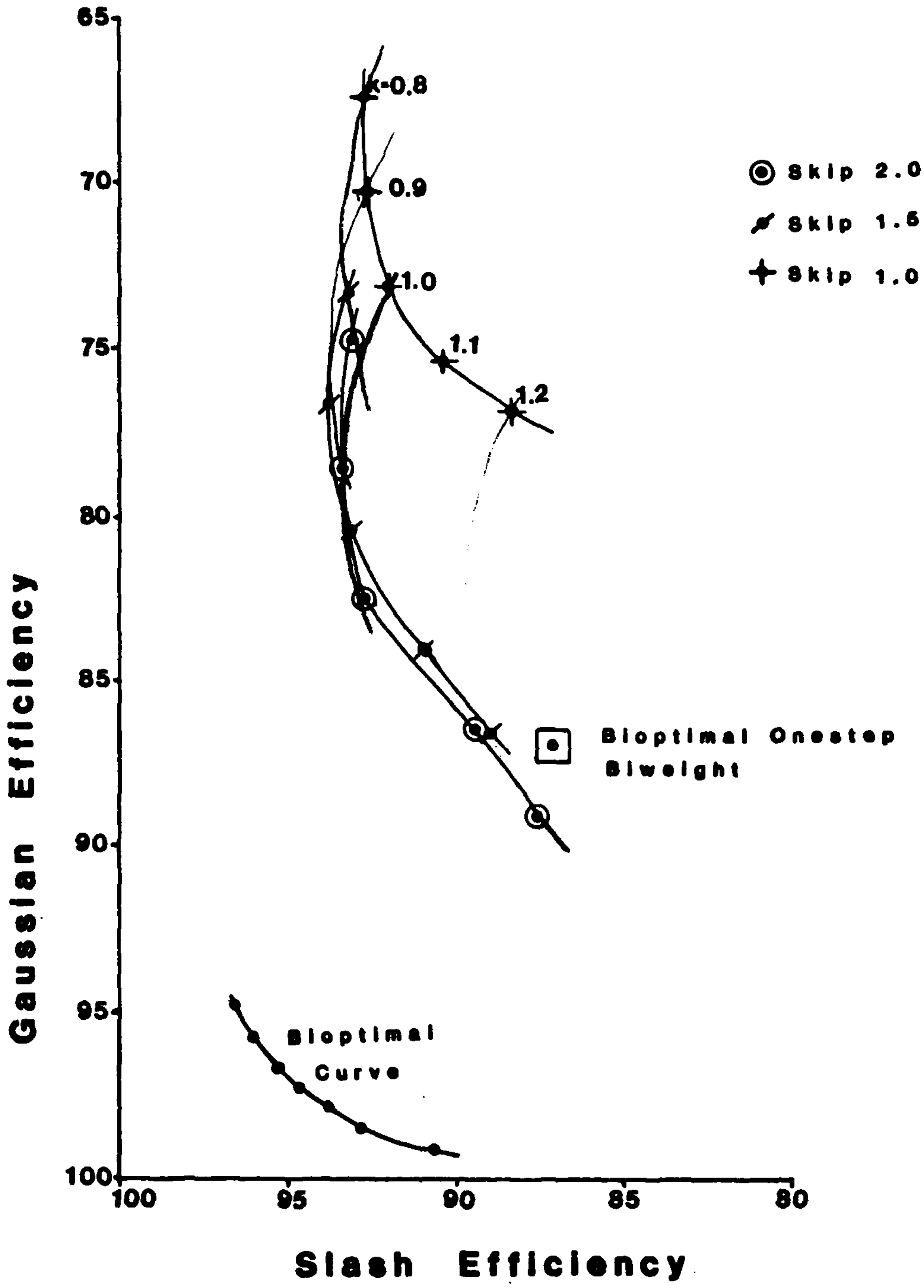

Figure 4: Bioptimal and skipped pushback curves.

efficiency, and 3) high slash efficiency with 95% Gaussian efficiency? The curve for a specified skipping factor roughly moves down on the efficiency/efficiency plot as the skipping factor increases. Thus one suggestion for increasing biefficiency is a pushback with skipping factor slightly larger than 2.0.

This suggestion may also be useful in achieving higher slash efficiency for 90% or 95% Gaussian efficiency. The slope on the right side of the fixed skipping factor curve increases with increasing skipping factor. Thus we would expect intersection with the 90% or 95% Gaussian efficiency horizontal line at a higher slash efficiency. The gains from increasing the skipping constant are not expected to be as large as those from the proposals below.

A second suggestion for increasing biefficiency and slash efficiency for 90% or 95% Gaussian efficiency is the set of estimates of the form

$$\theta \text{ skipped} + (1-\theta) \text{ unskipped} .$$

The pushback constants chosen for the skipped and unskipped versions used in the linear combination will depend on which of the aims 1)-3) is considered. Figure 4 indicates that for aims 2) and 3) larger pushback constants should be used than for aim 1).

Preliminary results on the performance of estimates of the form

θ skipped + $(1-\theta)$ unskipped

are given in figure 5. Figure 5 shows the skip at 2.0 push-back, the no skip pushback and the linear combination push-back efficiencies. For the linear combination pushback, the skipped pushback constant used is 1.2 and the unskipped pushback constant is 1.0. These results indicate that estimates of the linear combination form are likely to be a good choice for aims 1)-3).

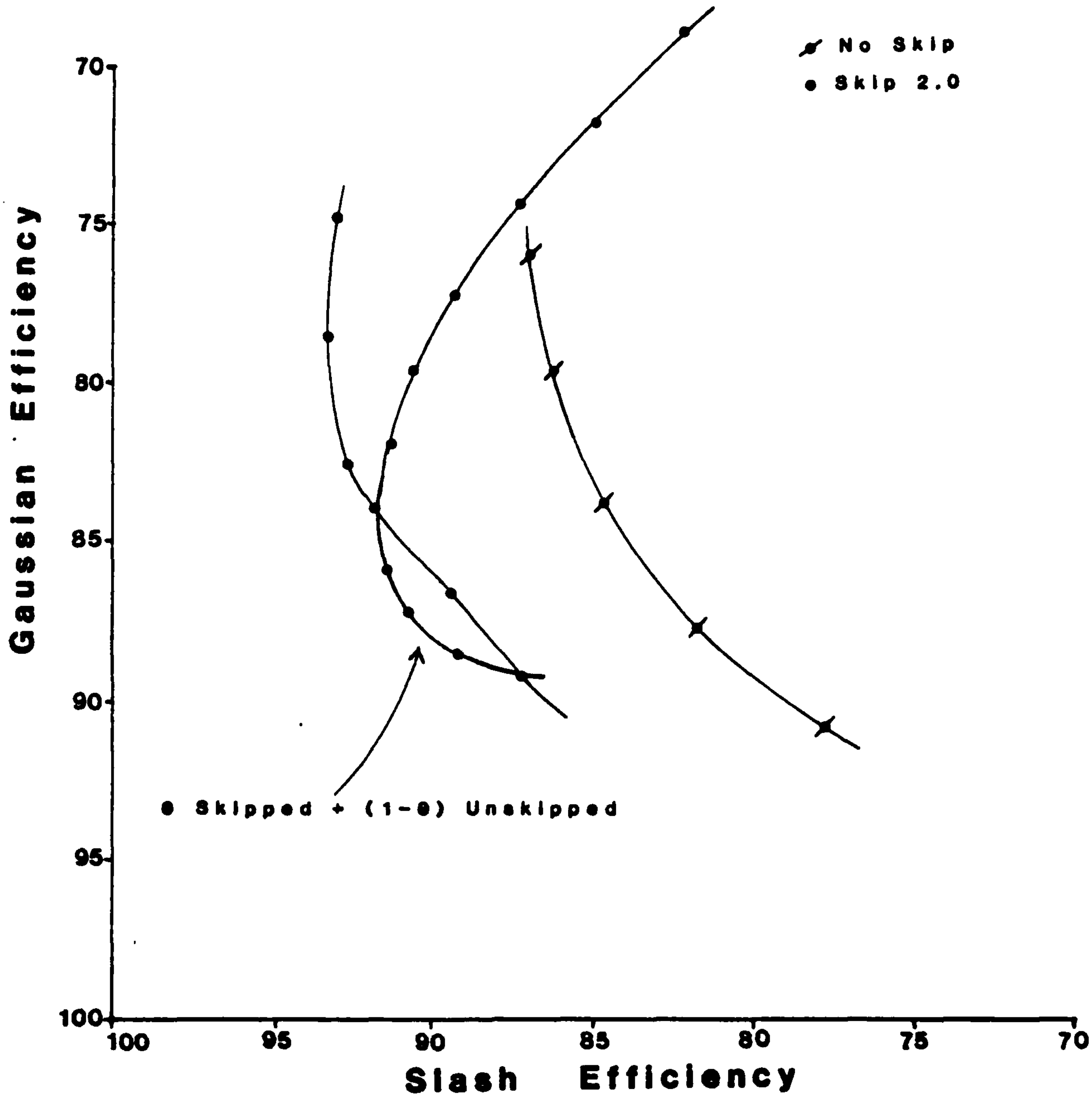

Figure 5: Skipped pushback, pushback, and linear combination curves.

REFERENCES

Andrews, D.F. et al (1972). Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, New Jersey.

Bell, K. and Pregibon, D. (1981). "Some computational details of configural sampling," Technical Report No. 191, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey.

Krystinik, K. B. (1981a). Data Modifications Based on Order: Pushback; A Configural Polysampling Approach, Ph.D. thesis, Department of Statistics, Princeton University, Princeton, New Jersey.

Krystinik, K. B. (1981b). "Pre-configural polysampling pushback performance," Technical Report No. 210, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey.

Krystinik, K. B. and Morgenthaler, S. (1981). "Comparison of the bioptimal curve with curves for two robust estimates," Technial Report No. 195, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey.

Pregibon, D. and Tukey, J.W. (1981). "Assessing the behavior of robust estimates of location in small samples: introduction to configural polysampling," Technical Report No. 185, Series 2, Department of Statistics, Princeton University, Princeton, N.J.

Tukey, J.W. (1981). "Some advanced thoughts on the data analysis involved in configural polysampling directed toward high performance estimates," Technical Report No. 189, Series 2, Department of Statistics, Princeton University, Princeton, New Jersey.